REZARTA ISLAMAJ DOĞAN

National Center for Biotechnology Information
8600 Rockville Pike 10N03D
Bethesda, MD 20894

(301) 435-8769
islamaj@ncbi.nlm.nih.gov

Research Interests

My research focuses on applying machine learning and data mining approaches for identifying useful information in biomedical databases. I work with PubMed log data in order to understand user needs and their search habits for improving biomedical information retrieval at NCBI. I am also interested in discovering and building domain appropriate features in order to model the biomedical information for accurate classification and prediction.

Education

UNIVERSITY OF MARYLAND — Ph.D. Computer Science
Ph.D. thesis: "Feature generation and analysis applied to sequence classification for splice-site prediction" — Fall 2007
Advisor: Lise Getoor

UNIVERSITY OF MARYLAND — M.S. Computer Science
M.S. scholarly paper: "Identification of protein-coding regions" — May 2003
Advisor: Lise Getoor

BOSPHORUS UNIVERSITY, ISTANBUL — B.Sc. Computer Engineering
B.S. thesis: "Three-dimensional representation of amino-acid characteristics" — June 2000
Advisors: Ethem Alpaydın and Uğur Sezerman
graduated with highest honors

Research and Professional Experience

RESEARCH FELLOW — National Center for Biotechnology Information
August, 2008 - present — Bethesda, MD
My research focuses on understanding user search habits of the PubMed database in order to improve biomedical information retrieval at NCBI. PubMed is crucial for researchers to keep abreast of the literature concerning their own research. As a first-of-its-kind work performed in such a large scale, I analyzed several months' worth of PubMed log data, consisting of millions of users' queries and their clicks. We identified and characterized unique aspects of PubMed searches and those results are directing and guiding other text-mining work on building new features for improving user search experiences in PubMed.

POST-DOCTORAL RESEARCH ASSOCIATE — Department of Cell Biology and Molecular Genetics
March, 2008 - July, 2008 — College Park, MD
In collaboration with Dr. Steve Mount and Dr. Lise Getoor I developed my sequence analysis model to detect biologically significant signals that effect splicing. Moreover, I designed a model that made use of the splice site prediction algorithm to identify sequence single-point mutations with a possible effect on the gene-splicing mechanism. These predicted "weak-point" mutations could be used by scientists to identify the exact points in the genetic sequence that have the most significant outcome, to target for in vivo experiments.

DISSERTATION RESEARCH
For the problem of sequence classification, I built an integrated process, which I referred to as *feature generation*. This algorithm allows the user to construct interesting features out of basic elements and to search effectively a large space of potential features. I applied this approach to the problem of splice-site prediction and achieved significant improvements in accuracy over existing, state-of-the-art approaches. I used the identified sets of features to discover biologically interesting motifs. For this, I created *SplicePort* (www.spliceport.org), an easy-to-use website that can be used to predict new splice sites from user-input sequences, and to browse the whole collection of features for informative signals. I also expanded the algorithm to construct more complex features, that also capture the three-dimensional

characteristics of the genomic sequence.

PRE-DOCTORAL FELLOW — NCBI/NLM/NIH
June, 2002 - December 2007 — Bethesda, MD
My research focused on pre-mRNA sequence analysis and splice-site prediction improvement through machine-learning approaches. I got familiar with NCBI tools and databases. I gained knowledge in algorithms and heuristics used in analyzing biological sequences, including computational gene finding, string matching, pattern finding, suffix and decision trees, alignment and dynamic programming algorithms.

TEACHING ASSISTANT — University of Maryland
September, 2000 - May, 2002 — College Park, MD
I assisted for Software Engineering and Database Design. I conducted discussion sessions and gave lectures, prepared assignments, prepared and graded term projects, graded exams and assisted students. The term projects for the Database Design course were based on the Oracle database.

TEACHING ASSISTANT — Bosphorus University
September, 1999 - June, 2000 — Istanbul, Turkey
I prepared the course outline and presented material for two terms of Introduction to Pascal and C programming courses. I conducted discussion sections and prepared assignments. I graded assignments, term projects and exams. I helped students individually during office hours and directed them in group work during lab hours. I organized a competition challenge for term projects.

INTERN Software Engineer — SuperOnline
June 1999 - September 1999 — Istanbul, Turkey
I implemented several parts of an on-going project using data-driven dynamic HTML generation with Sybase Power Dynamo over a Sybase database.

INTERN Application Developer — IDB
August 1998 - September 1998 — Istanbul, Turkey

INTERN Programmer — GigaByte
June 1998 - July 1998 — Istanbul, Turkey

Publications

Book Chapters

[1] "A feature generation algorithm with applications to biological sequence classification" with Lise Getoor and W. John Wilbur, Chapter in *Computational Methods of Feature Selection*, Huan Liu and Hiroshi Motoda editors (2007).

Journal Papers

[1] "Click-words: Learning to Predict Document Keywords from a User Perspective" with Zhiyong Lu *Bioinformatics* (August 2010)

[2] "Extracting Rx Information from Clinical Narrative" with James G. Mork, Olivier Bodenreider, Dina Demner-Fushman, Franois-Michel Lang, Zhiyong Lu, Aurlie Nvol, Lee Peters, Sonya E. Shooshan and Alan R. Aronson *JAMIA* (June 2010)

[3] "Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction" with Aurlie Nvol, and Zhiyong Lu *JBI* (submitted)

[4] "Understanding PubMed user search behavior through log analysis" with G. Craig Murray, Aurlie Nvol, and Zhiyong Lu *Database (Oxford)*, (November 2009)

[5] "Features generated for computational splice-site prediction correspond to functional elements" with Lise Getoor, W. John Wilbur and Stephen M. Mount, *BMC Bioinformatics* (October 2007).

[6] "SlicePort: an interactive splice-site analysis tool" with Lise Getoor, W. John Wilbur and Stephen M. Mount, *Nucleic Acids Research,* (June 2007).

[7] "Structural footprinting in protein structure comparison: the impact of structural fragments." with Elena Zotenko, W. John Wilbur, Diane P. O'Leary and Teresa M. Przytycka, *BMC Structural Biology,* (August 2007).

Refereed Conferences

[1] "A novel textual representation scheme for identifying clinical relationships in patients records" with Aurlie Nvol and Zhiyong Lu *ICMLA*, (December 2010)

[2] "Identifying protein-protein interactions in biomedical text articles" with Yi Yang, Aurlie Nvol, Minlie Huang and Zhiyong Lu *BioCreative III,* (September 2010)

[3] "Towards Efficient Search Tools for Biomedical Databases: Characterizing User Search Habits and Recognizing their Information Needs" with G. Craig Murray, Aurlie Nvol, and Zhiyong Lu *ISMB* (July 2010)

[4] "Characterizing User Search Behavior in PubMed" with G. Craig Murray, Aurlie Nvol, and Zhiyong Lu *AMIA* (November 2010)

[5] "Author keywords in Biomedical Journal Articles" with Aurlie Nvol, and Zhiyong Lu ÂMIA (November 2010)

[6] "Visualizing the weakest links: nucleotides vulnerable to mutations that affect splicing" with Stephen M. Mount and Lise Getoor, *Alternative Splicing-Special Interest Group Meeting at ISMB 2008,* Toronto, Canada (July 2008).

[7] "Characterizing RNA secondary-structure features and their effects on splice-site prediction" with Lise Getoor and W. John Wilbur, *IEEE ICDM Workshop on Mining and Management of Biological Data,* (October 2007).

[8] "Feature generation algorithm: with application to splice-site prediction," with Lise Getoor and W. John Wilbur, *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases,* Berlin, Germany (September 2006).

[9] "A feature generation algorithm for sequences with application to splice-site prediction" with Lise Getoor and W. John Wilbur, *International Workshop on Feature Selection for Data Mining: Interfacing Machine Learning and Statistics,* Bethesda, Maryland (April 2006).

[10] "Three dimensional representation of amino acid characteristics" with Ugur Sezerman and Ethem Alpaydin, *23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society,* Istanbul, Turkey (October 2001).

Talks and Presentations

[1] "A hybrid approach for PPI content identification and method detection" *BioCreative III*, Bethesda, MD (September 2010)

[2] "Identifying words that users find important for retrieving relevant MEDLINE articles" *NCBI Computational Biology Branch Seminar*, Bethesda, MD (January 2010)

[3] "Machine Learning Techniques for Biomedical Text Retrieval in PubMed" (**invited tutorial**) with Lana Yeganova *International Conference on Machine Learning and Applications,* Miami, FLA (December 2009)

[4] "From recognizing biological sequences, to identifying search keywords: A feature generation framework" (**invited talk**) *GRAND Seminar, George Mason University,* Fairfax, VA (September 2009)

[5] "Visualizing the weakest links: nucleotides vulnerable to mutations that affect splicing" with Stephen M. Mount and Lise Getoor, *Alternative Splicing-Special Interest Group Meeting at ISMB 2008,* Toronto, Canada (July 2008).

[6] "Feature analysis for splice-site prediction" (**invited talk**) *Department of Biology, Massachusetts Institute of Technology,* Boston, MA (May 2008).

[7] "Feature analysis for splice-site prediction" (**invited talk**) *Penn Genomics Institute, University of Pennsylvania,* Philadelphia, PA (December 2007).

[8] "Feature generation analysis and splice-site prediction" (**invited talk**) *Memorial Sloan-Kettering Cancer Center,* New York, NY (November 2007).

[9] "Analysis of splicing motifs" (**invited talk**) *Splicing Regulator Motifs Workshop,* Erlangen, Germany (September 2006).

[10] "The proximity effect on feature construction for splice-site finding" *NIH Graduate Student Research Symposium,* Bethesda, Maryland (May 2006).

[11] "A machine learning solution for splice-site prediction" *TASSA Annual Conference,* Philadelphia, PA (March 2006).

[12] "Feature generation for sequences with application to splice-site prediction" *NIH Research Festival,* Bethesda, Maryland (October 2005).

[13] "Finding acceptor splice sites with AdaBoost" *NIH Second Annual Graduate Student Research Symposium,* Bethesda, Maryland (April 2005).

[14] "Classification of acceptor splice sites using AdaBoost with decision trees" *16th Annual Genomic Sequencing and Analysis Conference,* Washington DC (September 2004).

[15] "Identification of internal coding regions in mRNA sequences" *NIH First Annual Graduate Student Research Symposium,* Bethesda, Maryland (April 2004).

Skills

TECHNICAL SKILLS: Windows, UNIX/Linux, Macintosh OS / OS X. Applications - LaTeX, MS Office, Emacs, etc.
PROGRAMMING: C, C++, CGI, Matlab, SQL, HTML, XML, Java, Perl, etc.
DATABASES: Oracle, Sybase, MS-SQL, MS-Access.
BIOINFORMATICS: Knowledgeable in a variety of bioinformatics tools and databases, including but not limited to: Blast, PSI-Blast, Pubmed, ClustalW, GeneSplicer, MaxEnt, RescueESE, SplicePredictor, HMMGene, NetGene, GenScan, ExonScan, Muscle, SplicePort.
LANGUAGES: English, Turkish, Albanian – fluent; Italian, French – good; Spanish, German – beginner.

Honors and Awards

NIH pre-doctoral Fellow	2002-2007
Faculty Horizons	June 2006
Graduated 5th, Computer Science, Bosphorus(Bogazici) University	June 2000
Dean's Honor List	1996-2000
Turkish Prime Ministry Fellowship	1997-2000
Rumeli Foundation Fellowship	1997-2000
Ranked 1st in National Mathematical Olimpiad, Albania	March 1996
Honorable Mention in 13th Balkan Mathematical Olympiad, Romania	April 1996
37th International Mathematical Olympiad, India	July 1996